PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Bayesian inference and comparison of stochastic transcription elongation models

Jordan Douglas[1,2]*, **Richard Kingston**[1], **Alexei J. Drummond**[1,2]

1 School of Biological Sciences, University of Auckland, Auckland, New Zealand, **2** Centre for Computational Evolution, School of Computer Science, University of Auckland, Auckland, New Zealand

* jordan.douglas@auckland.ac.nz

Abstract

Transcription elongation can be modelled as a three step process, involving polymerase translocation, NTP binding, and nucleotide incorporation into the nascent mRNA. This cycle of events can be simulated at the single-molecule level as a continuous-time Markov process using parameters derived from single-molecule experiments. Previously developed models differ in the way they are parameterised, and in their incorporation of partial equilibrium approximations. We have formulated a hierarchical network comprised of 12 sequence-dependent transcription elongation models. The simplest model has two parameters and assumes that both translocation and NTP binding can be modelled as equilibrium processes. The most complex model has six parameters makes no partial equilibrium assumptions. We systematically compared the ability of these models to explain published force-velocity data, using approximate Bayesian computation. This analysis was performed using data for the RNA polymerase complexes of *E. coli*, *S. cerevisiae* and Bacteriophage T7. Our analysis indicates that the polymerases differ significantly in their translocation rates, with the rates in T7 pol being fast compared to *E. coli* RNAP and *S. cerevisiae* pol II. Different models are applicable in different cases. We also show that all three RNA polymerases have an energetic preference for the posttranslocated state over the pretranslocated state. A Bayesian inference and model selection framework, like the one presented in this publication, should be routinely applicable to the interrogation of single-molecule datasets.

Author summary

Transcription is a critical biological process which occurs in all living organisms. It involves copying the organism's genetic material into messenger RNA (mRNA) which directs protein synthesis on the ribosome. Transcription is performed by RNA polymerases which have been extensively studied using both ensemble and single-molecule techniques. Single-molecule data provides unique insights into the molecular behaviour of RNA polymerases. Transcription at the single-molecule level can be computationally simulated as a continuous-time Markov process and the model outputs compared with experimental data. In this study we use Bayesian techniques to perform a systematic comparison of 12 stochastic models of transcriptional elongation. We demonstrate how

Citation: Douglas J, Kingston R, Drummond AJ (2020) Bayesian inference and comparison of stochastic transcription elongation models. PLoS Comput Biol 16(2): e1006717. https://doi.org/10.1371/journal.pcbi.1006717

Editor: Vassily Hatzimanikatis, Ecole Polytechnique Fédérale de Lausanne, SWITZERLAND

Received: December 12, 2018

Accepted: December 13, 2019

Published: February 14, 2020

Data Availability Statement: All posterior distribution files and the software used to generate these distributions are available in a GitHub repository https://github.com/jordandouglas/jordandouglas.github.io/tree/master/simpol/about/ElongationPosteriorData.

Funding: J. Douglas was funded by a University of Auckland Doctoral Scholarship. URL: https://www.auckland.ac.nz/en/study/scholarships-and-awards/scholarship-types/postgraduate-scholarships/doctoral-scholarships.html. The authors wish to acknowledge the contribution of NeSI highperformance computing facilities to the results of this research. NZ's national facilities are provided by the NZ eScience Infrastructure and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme. URL https://www.nesi.org.nz. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

equilibrium approximations can strengthen or weaken the model, and show how Bayesian techniques can identify necessary or unnecessary model parameters. We describe a framework to a) simulate, b) perform inference on, and c) compare models of transcription elongation.

Introduction

Transcription is carried out by RNA polymerases: RNAP in *Escherichia coli*, pol II in *Saccharomyces cerevisiae*, and T7 pol in Bacteriophage T7. It involves the copying of template double-stranded DNA (dsDNA) into single-stranded messenger RNA (mRNA). RNAP and pol II are comprised of multiple subunits, and their catalytic subunits are homologous [1, 2]. In contrast, T7 pol exists as a monomer with a distinct sequence, and resembles the *E. coli* DNA polymerase I [3].

Optical trapping experiments have been performed on the transcription elongation complex (TEC) from a variety of organisms [4–10]. In a typical experimental setup, two polystyrene beads (around 600 nm in diameter) are tethered to the system; one attached to the RNA polymerase and the other to the DNA [4]. As transcription elongation progresses, the distance between the two beads increases and the velocity of a single TEC can be computed. Optical tweezers can be used to apply a force *F* to the system (Fig 1).

Single-molecule studies of the TEC have revealed that RNA polymerases progress in a discontinuous fashion [4, 11–14] with step sizes that correspond to the dimensions of a single nucleotide (3.4 Å [15]). Consequently, at the single molecule level, transcription is best modelled as a discrete process rather than a continuous one.

A single cycle in the main transcription elongation pathway (Fig 2) requires (1) Forward translocation of the RNA polymerase, making the active site accessible; (2) Binding of the complementary nucleoside triphosphate (NTP); (3) Addition of the nucleotide onto the 3′ end of the mRNA. This third step involves NTP hydrolysis. Nucleoside monophosphate is added onto the chain and pyrophosphate is released from the enzyme.

Our study aimed to identify the best model to describe this reaction cycle for RNAP, pol II, and T7 pol, based on analysis of published force-velocity data. As there are three reactions, up to six rate constants may be necessary for a kinetic model of a single nucleotide addition. These describe forward and backwards translocation (k_{fwd} and k_{bck}), binding and release of NTP (k_{bind} and k_{rel}), and NTP catalysis and reverse-catalysis (k_{cat} and k_{rev}), also known as pyrophosphorolysis [18]. However fewer than six parameters may be required in practice.

First, it is reasonable to assume that polymerisation is effectively irreversible [17, 19–21], as pyrophosphorolysis is a highly exergonic reaction, reducing the number of rate constants to five. Second, translocation between the pretranslocated and posttranslocated states, and/or NTP binding, may occur on timescales significantly more rapid than the other steps, in which case they may be modelled as equilibrium processes. These assumptions simplify the model, as the respective forward and reverse reaction rate constants are subsumed by a single equilibrium constant. Third, thermodynamic models of nucleic acid structure can be used to estimate sequence-dependent translocation rates $k_{fwd}(l)$ and $k_{bck}(l)$, by invoking transition state theory, and this can sometimes result in parameter reduction [16, 17, 21].

Irrespective of equilibrium assumptions and parameterisation, transcription elongation under applied force can be modelled in two fundamentally distinct ways. First, there are the **deterministic** equations which can be used to calculate the mean pause-free elongation

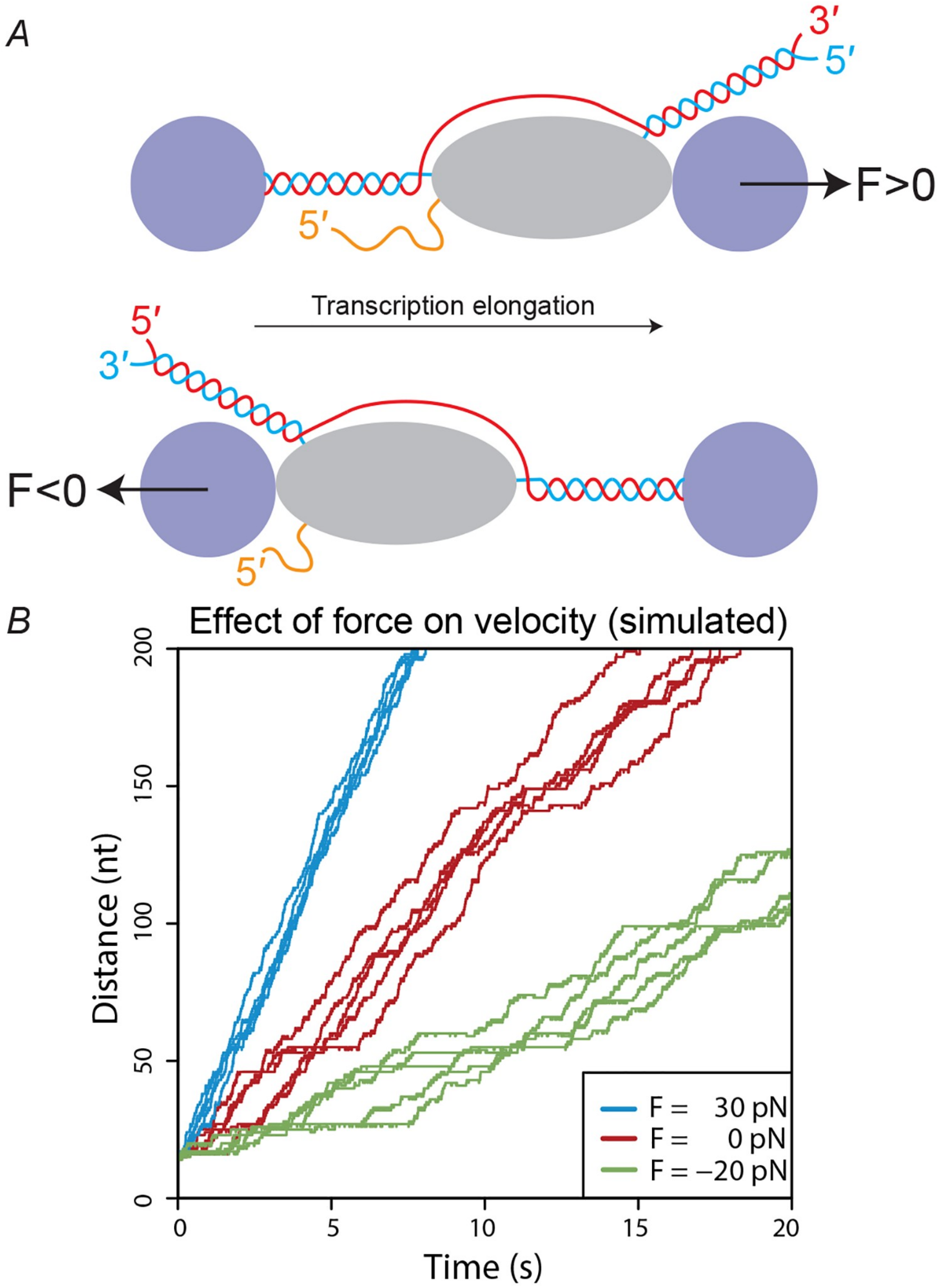

Fig 1. Effect of an applied force on elongation velocity. (A) Optical trapping setup showing dsDNA being transcribed by RNA polymerase (grey ellipse) into mRNA. Two polystyrene beads are tethered to the system allowing the application of force using optical tweezers. An assisting load $F > 0$ acts in the same direction as transcription (top) while a hindering load $F < 0$ acts in the opposing direction (bottom). Figure not to scale. (B) Schematic depiction of the effect of applying a force on RNA polymerase. Due to the stochastic nature of transcription at the single-molecule level, each experiment yields a different distance-time trajectory, even under the same applied force.

https://doi.org/10.1371/journal.pcbi.1006717.g001

velocity $v(F, [\mathrm{NTP}])$ as a function of force F and NTP concentration [NTP]. This kind of model can be derived from the differential equations describing the time evolution of all species, by application of the steady state approximation. Force effects on the translocation step are incorporated using transition state theory [22, 23].

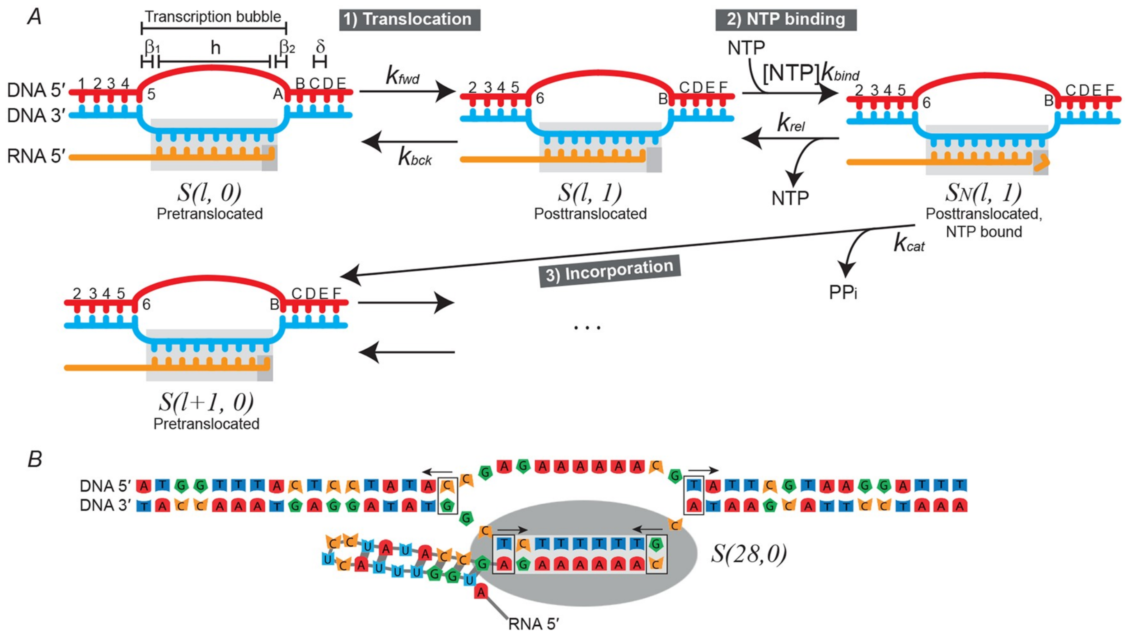

Fig 2. State diagrams of RNA polymerase. (A) The model of the main transcription elongation pathway, which shows the postulated states; the pathways for interconversion; and the rate constants that govern each part of the reaction. The transcription bubble is the set of $\beta_1 + h + \beta_2$ bases (see main text for definitions) in the double-stranded DNA which are unpaired. States are denoted by $S(l, t)$ where l is the length of the mRNA and t is the position of the polymerase active site (small grey rectangle) with respect to the 3′ end of the mRNA. Polymerase translocation displaces the polymerase by a distance of δ = 1 bp = 3.4 Å. During polymerisation the chain is extended by one nucleotide. (B) Instantiated posttranslocated state of RNA polymerase transcribing the *rpoB* gene sequence, with $\beta_1 = 2$, $h = 9$, $\beta_2 = 1$. Forward translocation requires melting two T/A basepairs (right arrows). Backward translocation requires melting two C/G basepairs (left arrows). The mRNA secondary structure would also require reconfiguration [16, 17].

https://doi.org/10.1371/journal.pcbi.1006717.g002

An example is the following 3-parameter model [4].

$$v(F, [\mathrm{NTP}]) = \frac{k_{cat}}{1 + \frac{K_D}{[\mathrm{NTP}]}\left(1 + K_\tau e^{-F\delta/k_B T}\right)} \tag{1}$$

where δ is the distance between adjacent basepairs (3.4 Å, [15]), $K_D = \frac{k_{rel}}{k_{bind}}$ is the equilibrium constant of NTP binding, $K_\tau = \frac{k_{bck}}{k_{fwd}}$ is the equilibrium constant of translocation, k_B is the Boltzmann constant, and T is the absolute temperature. Increasingly complex equations may be used as more parameters or states are added to the model [4, 6, 17]. Such equations describe the velocity averaged across an ensemble of molecules. Parameter inference applied to velocity-force-[NTP] experimental data is straightforward and computationally fast when using these equations. However these equations do not describe the distribution of velocity nor do they account for site heterogeneity across the nucleic acid sequence and therefore cannot predict local sequence effects.

Second, there are the **stochastic** models, which can be implemented via simulation of single-molecule behaviour using the Gillespie algorithm [24]. The mean velocity can be calculated by averaging velocities over a number of simulations for a given F and [NTP]. This offers not just the mean but a full distribution of velocities and could potentially explain emergent properties unavailable from a deterministic model. Unfortunately, simulating can be very slow and therefore parameter inference can be a problem.

In this study we used a Markov-chain-Monte-Carlo approximate-Bayesian-computation (MCMC-ABC) algorithm [25] to estimate transcription elongation parameters for **stochastic** models via simulation. The observed pause-free velocities we are fitting to were measured at varying applied force and NTP concentration. For each RNA polymerase under study—*E. coli* RNAP, *S. cerevisiae* pol II, and T7 pol—we fit to one respective dataset from the single-molecule literature [4, 26, 27].

Models

Notation and state space

Suppose the TEC is transcribing a gene of length L. Then let $S(l, t)$ denote a TEC state, where the mRNA is currently of length $l \leq L$, and $t \in \mathbb{Z}$ describes the position of the active site with respect to the 3′ end of the mRNA. When $t = 0$ the polymerase is pretranslocated and cannot bind NTP, and when $t = 1$ the polymerase is posttranslocated and *can* bind NTP (Fig 2). This study is focused on the main elongation pathway and the observed velocities being fitted have pauses filtered out. Therefore, although additional backtracked states ($t < 0$) [4, 28, 29] and hypertranslocated states ($t > 1$) [30, 31] exist, these are not incorporated in the model.

Let β_1 and β_2 be the number of unpaired template nucleotides upstream and downstream of RNA polymerase, respectively, and let h be the number of basepairs in the DNA/mRNA hybrid (Fig 2A). Although there are uncertainties in these parameters, they are held constant at $h = 9$, $\beta_1 = 2$, and $\beta_2 = 1$ [17, 32].

Transcription of the gene begins at state $S(l_0, 0)$ and ends upon reaching $S(L, 0)$, where $l_0 = \beta_1 + h + 2$.

Parameterisation of the NTP binding step

NTP binding has been modelled as both a kinetic and equilibrium process in the literature [4, 17, 21].

In a kinetic binding model, NTP binding occurs at pseudo-first order rate k_{bind}[NTP], while NTP release occurs at rate k_{rel}. In this case, k_{bind} and $\frac{k_{rel}}{k_{bind}}$ must be estimated.

Under a partial equilibrium approximation NTP binding and release are assumed to be rapid enough that equilibrium is achieved. In this case, the rate constants k_{bind} and k_{rel} are subsumed by the NTP dissociation constant $K_D = \frac{k_{rel}}{k_{bind}}$ which becomes the sole binding-related parameter to estimate.

Parameterisation of the translocation step

While inferences about the rate constants associated with NTP binding and catalysis (k_{bind}, $\frac{k_{rel}}{k_{bind}}$, and k_{cat}) can be made directly from the data, the translocation step is more complex. Transition state theory is invoked in order to estimate k_{fwd} and k_{bck}. Recasting the problem in this way (1) provides a way of accommodating the effects of applied force on the elongation process, and (2) allows the sequence-dependence of translocation to be incorporated by considering the energetics of basepairing. When allowing for sequence dependence, the total number of translocation rates required to model translocation of the full gene is $2(L - l_0)$.

Thermodynamic models of base pairing energies. The standard Gibbs free energies $\Delta_r G^0 (= \Delta G)$ involved in duplex formation are calculated using nearest neighbour models. The standard Gibbs energy of state S—arising from nucleotide basepairing and dangling ends—is calculated as

$$\Delta G_S^{(bp)} = \Delta G_{gene}^{(bp)} + \Delta G_{hybrid}^{(bp)} \tag{2}$$

where SantaLucia's DNA/DNA basepairing parameters [33] are used to calculate $\Delta G_{gene}^{(bp)}$ and Sugimoto's DNA/RNA parameters [34] are used for $\Delta G_{hybrid}^{(bp)}$. For the latter, dangling end energies are estimated as described by Bai et al. 2004 [21]. Here, and elsewhere, the $^{(bp)}$ superscript is used to denote a model parameter that can be evaluated from the sequence alone. Gibbs energies are expressed on a per molecule basis, relative to the thermal energy of the system, in multiples of $k_B T$, where $k_B T = 4.28001 \times 10^{-21}$ J at $T = 310$ K.

In order for RNA polymerase to translocate forward (backward), up to two basepairs must be disrupted: (1) the basepair at the downstream (upstream) edge of the transcription bubble, and (2) the basepair at the upstream (downstream) end of the DNA/mRNA hybrid (Fig 2B). Differences in the basepairing energies in these regions confer sequence-dependence on the rate of translocation.

Calculation of translocation rates or translocation equilibrium constant. The standard Gibbs energies of the pre and posttranslocated states, $\Delta G_{S(l,0)}^{(bp)}$ and $\Delta G_{S(l,1)}^{(bp)}$, respectively, are used with up to four additional terms—$\Delta G_{\tau 1}$, δ_1, $\Delta G_\tau^\ddagger$, and $\Delta G_{T(l,t)}^{(bp)}$—to calculate the translocation rates. The first three are model parameters which must be estimated while the latter is directly evaluated from the sequence.

Let $T(l, t)$ be the translocation transition state between $S(l, t)$ and $S(l, t + 1)$. Then $\Delta G_{T(l,t)}^\ddagger = \Delta G_\tau^\ddagger + \Delta G_{T(l,t)}^{(bp)}$ is the sequence-dependent standard Gibbs energy of activation which must be overcome in order to translocate (Fig 3).

Given an applied force F, the translocation rates governing transition between the pre and posttranslocated states ($k_{fwd}(l)$ and $k_{bck}(l)$) are calculated from barrier height $\Delta G_{T(l,0)}^\ddagger$ using an Arrhenius type relation:

$$k_{fwd}(l) = Ae^{-(\Delta G_{T(l,0)}^\ddagger - \Delta G_{S(l,0)}^{(bp)} - F\delta_1 / k_B T)} \tag{3}$$

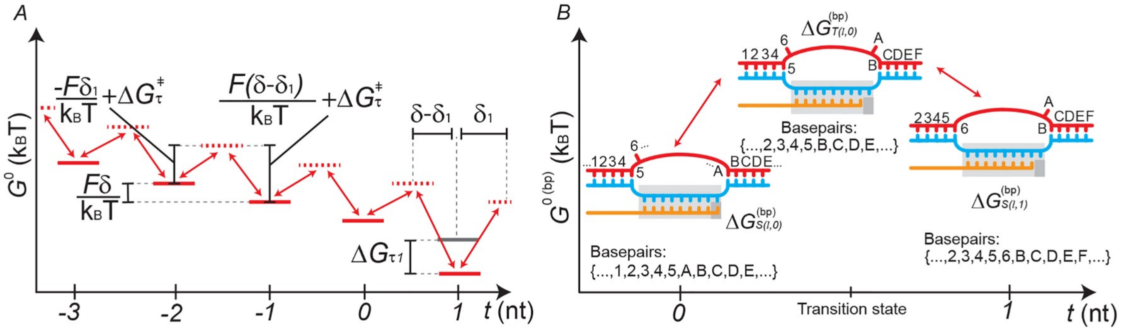
A
G^0 ($k_B T$)
$\frac{-F\delta_1}{k_B T} + \Delta G_\tau^\ddagger$
$\frac{F(\delta-\delta_1)}{k_B T} + \Delta G_\tau^\ddagger$
$\frac{F\delta}{k_B T}$
$\delta-\delta_1$
δ_1
$\Delta G_{\tau 1}$
-3
-2
-1
0
1
t (nt)
B
$G^{0\,(bp)}$ ($k_B T$)
$\Delta G_{T(l,0)}^{(bp)}$
1234
6
5
A
B
CDEF
Basepairs:
{...,2,3,4,5,B,C,D,E,...}
...1234
6...
5
...A
BCDE...
$\Delta G_{S(l,0)}^{(bp)}$
Basepairs:
{...,1,2,3,4,5,A,B,C,D,E,...}
2345
6
A
B
CDEF
$\Delta G_{S(l,1)}^{(bp)}$
Basepairs:
{...,2,3,4,5,6,B,C,D,E,F,...}
0
Transition state
1
t (nt)

Fig 3. Parameterisation of the translocation step. (A) Effects of model parameters on state energies. The figure displays a schematic Gibbs energy landscape of translocation, with backtracked states included for visualisation purposes. The solid red lines represent translocation states ($t = 0$: pretranslocated, $t = 1$: posttranslocated, and $t < 0$: backtracked), while the dashed red lines represent transition states. Applying an assisting force $F > 0$ tilts the landscape in favour of higher values of t. The effect of $\Delta G_{\tau 1}$ is observed at the posttranslocated state $t = 1$. In a translocation equilibrium model, the barrier height is assumed to be so small, = translocation is so rapid, that the transition states are disregarded. (B) A model for the sequence-dependent transition state between translocation states $S(l, 0)$ and $S(l, 1)$. This is required for estimating the Gibbs energy of basepairing $\Delta G_{T(l,t)}^{(bp)}$ in the transition state. The basepairing energy, added to a baseline term $\Delta G_\tau^\ddagger$, together specify the height of the activation barrier (Eq 10).

https://doi.org/10.1371/journal.pcbi.1006717.g003

$$k_{bck}(l) = Ae^{-(\Delta G^{\ddagger}_{T(l,0)} - (\Delta G^{(bp)}_{S(l,1)} + \Delta G_{\tau 1}) + F(\delta - \delta_1)/k_B T)} \tag{4}$$

The derived rates $k_{fwd}(l)$ and $k_{bck}(l)$ are therefore dependent on the local sequence. The pre-exponential factor A is held constant at 10^6 s^{-1}. This term has been arbitrarily set to a variety of values in previous studies (10^6–10^9 s^{-1} [16, 17, 21]). This has little consequence for model fitting, however the value of $\Delta G^{\ddagger}_{T(l,t)}$ is entangled with the value of the pre-exponential factor A and can only be meaningfully interpreted in light of its value.

If the system has time to reach equilibrium, the probabilities of observing the pretranslocated state $S(l, 0)$ and posttranslocated state $S(l, 1)$ are

$$p(S(l,0)) \propto e^{-(\Delta G^{(bp)}_{S(l,0)})} \tag{5}$$

$$p(S(l,1)) \propto e^{-(\Delta G_{\tau 1} + \Delta G^{(bp)}_{S(l,1)})} \tag{6}$$

This is described by equilibrium constant K_τ.

$$K_\tau(l) \quad = \frac{p(S(l,0))}{p(S(l,1))} \tag{7}$$

$$= exp\{-(\Delta G_{S(l,0)} - \Delta G_{S(l,1)})\} \tag{8}$$

$$= exp\{-(\Delta G^{(bp)}_{S(l,0)} - \Delta G^{(bp)}_{S(l,1)} - \Delta G_{\tau 1})\} \tag{9}$$

The physical meanings of the terms $\Delta G_{\tau 1}$, δ_1, $\Delta G^{\ddagger}_{\tau}$, and $\Delta G^{(bp)}_{T(l,t)}$, and the way they are used in the model, are detailed below.

Energetic bias for the posttranslocated states. $\Delta G_{\tau 1}$ (units $k_B T$) is a parameter added to the standard Gibbs energy of the posttranslocated state. If $\Delta G_{\tau 1} = 0$, then the sequence alone determines the Gibbs energy difference between pre and posttranslocated states. In this case, pretranslocated states are usually favoured over posttranslocated states due to the loss of a single basepair in the hybrid of the latter.

$\Delta G_{\tau 1}$ has frequently been estimated for T7 pol [35–37] and there has been discussion around whether such a term is necessary for RNAP [6].

Polymerase displacement and formation of the transition state. δ_1 (units Å) is the distance that the polymerase must translocate forward to facilitate the formation of the transition state. The distance between adjacent basepairs is held constant at an experimentally measured value $\delta = 3.4$ Å [15], and $0 < \delta_1 < \delta$. The response of the system to an applied force F depends on this term. In general, the application of force F tilts the Gibbs energy landscape—the Gibbs energy difference between adjacent translocation states being augmented by a factor $\frac{F\delta}{k_B T}$ (Fig 3A, [38, 39]).

It may be necessary to estimate δ_1 to model the data adequately [17], or it may be sufficient to simply set $\delta_1 = \delta/2$ [38].

Energy barrier of translocation. $\Delta G^{\ddagger}_{\tau}$ and $\Delta G^{(bp)}_{T(l,t)}$ (units $k_B T$) together determine the activation barrier height in the translocation step. It is assumed that the sequence-dependent standard Gibbs energy of activation $\Delta G^{\ddagger}_{T(l,t)}$ can be written as

$$\Delta G^{\ddagger}_{T(l,t)} = \Delta G^{\ddagger}_{\tau} + \Delta G^{(bp)}_{T(l,t)} \tag{10}$$

$\Delta G_{\tau}^{\ddagger}$ is therefore a sequence-independent baseline term used to compute the translocation barrier heights. The parameter $\Delta G_{\tau}^{\ddagger}$ must be estimated in order to evaluate translocation rates.

In contrast $\Delta G_{T(l,t)}^{(bp)}$ is a term that is evaluated directly from the sequence derived from a model of the transition state (Fig 3B). The term is evaluated as the standard Gibbs energy of a TEC containing all hybrid and gene basepairs found in both $S(l, t)$ and $S(l, t + 1)$, ie. the intersection between the two sets of basepairs.

Model space

The full transcription elongation model makes use of the following 6 parameters:

- k_{cat} (units s^{-1}).
- $K_D = \frac{k_{rel}}{k_{bind}}$ (units μM).
- k_{bind} (units μM^{-1} s^{-1}).
- $\Delta G_{\tau 1}$ (units $k_B T$).
- δ_1 (units Å).
- $\Delta G_{\tau}^{\ddagger}$ (units $k_B T$).

However fewer than 6 parameters may be needed to adequately describe the data. If it is assumed that the energy differences between pre and posttranslocated states are determined by basepairing energies alone, the parameter $\Delta G_{\tau 1}$ does not need to be estimated. This is equivalent to holding $\Delta G_{\tau 1}$ constant at 0. If it is assumed that the displacement required for formation of the translocation transition state is half the distance between adjacent basepairs, the parameter δ_1 does not need to be estimated. This is equivalent to holding δ_1 constant at $\delta/2$.

Partial equilibrium approximations may also simplify the model, as detailed above. If binding is approximated as an equilibrium process, k_{bind} does not need to be estimated. If translocation is approximated as an equilibrium process, $\Delta G_{\tau}^{\ddagger}$ and δ_1 do not need to be estimated. One, both, or neither of these two steps (binding and translocation) could be assumed to achieve equilibrium, thus yielding four equilibrium model variants (Fig 4A). The introduction of partial equilibrium approximations for both the NTP binding and translocation steps has implications when specifying the prior distributions for the Bayesian analysis (S4 Appendix) The chemical master equations for single nucleotide addition cycles of these models are presented in S2 Appendix.

Incorporating these simplifications to the model in a combinatorial fashion results in a total of 12 related models, which together constitute the model space. Our objective was to determine which of these 12 models provides the best description of the experimental data. The simplest model (Model 1) contains 2 parameters (k_{cat} and K_D). The most complex model (Model 12) contains all 6 parameters. The full model space is displayed in Fig 4B.

Stochastic modelling

For each model we performed stochastic simulations, appropriate for the modelling of single-molecule force-velocity data. The simulations, performed using the Gillespie algorithm [24, 40], can be used to estimate the mean elongation velocity under a model.

The estimation of mean velocity can be broken down into three steps. First, the system is initialised by placing the RNA polymerase at the 3′ end of the template—state $S(l_0, 0)$—with the transcription bubble open and a DNA/RNA hybrid formed. The force and NTP concentrations are assigned their experimentally set values. Second, a chemical reaction is randomly

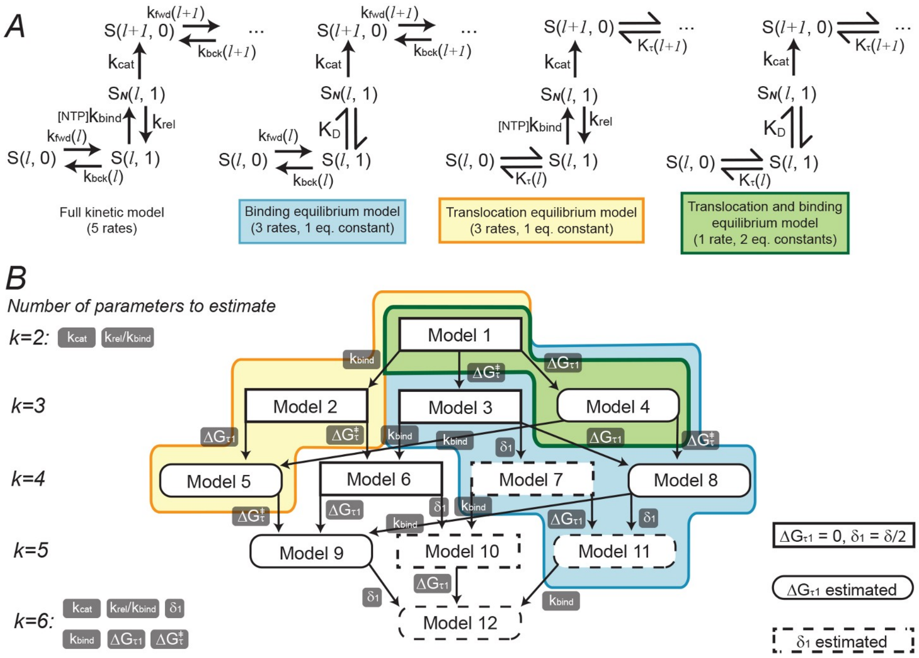
A
Full kinetic model (5 rates)
Binding equilibrium model (3 rates, 1 eq. constant)
Translocation equilibrium model (3 rates, 1 eq. constant)
Translocation and binding equilibrium model (1 rate, 2 eq. constants)
B
Number of parameters to estimate
k=2:
k=3
k=4
k=5
k=6:
Model 1
Model 2
Model 3
Model 4
Model 5
Model 6
Model 7
Model 8
Model 9
Model 10
Model 11
Model 12
ΔGτ1 = 0, δ1 = δ/2
ΔGτ1 estimated
δ1 estimated

Fig 4. The space of models to be compared. (A) The four equilibrium model variants. NTP binding, translocation, both, or neither, could be assumed to achieve equilibrium prior to catalysis. (B) The 12 transcription elongation models. An arrow connects model *i* to *j* if augmentation of model *i* with a single parameter generates model *j*. The number of parameters to estimate *k* is shown for each level in the network. Equilibrium approximation colour scheme is the same as in A. $\Delta G_{\tau 1}$ and δ_1 can each be estimated or set to a constant.

https://doi.org/10.1371/journal.pcbi.1006717.g004

sampled. The probability that reaction $S \xrightarrow{k} S'$ is selected is proportional to its rate constant k (Fig 2). The amount of time taken for the reaction to occur is sampled from the exponential distribution. States which are subject to a partial equilibrium approximation are coalesced into a single state, which augments the outbound rate constants. The second step is repeated until the RNA polymerase has copied the entire template. Third, the previous two steps are repeated c times. The mean elongation velocity is evaluated as the mean of each mean elongation velocity across c simulations. For further information, see S1 Appendix.

Relation to previous models and stochastic simulations

There is an extensive literature concerned with the kinetic modelling of transcription elongation. Such models may incorporate backtracking, hypertranslocation, and other reactions. Here we are concerned only with the central elongation pathway.

A stochastic and sequence-dependent model was proposed by Bai et al. 2004 [21] for RNAP, with both NTP binding and translocation treated as equilibrium processes. The translocation equilibrium constant was calculated entirely from basepairing energies. Therefore this model is equivalent to Model 1, and the parameters were estimated as $k_{cat} = 24.7$ s^{-1} and $K_D = 15.6$ μM from fit to experimental data. Maoiléidigh et al. 2011 also presented stochastic simulations of RNAP. The elongation component of their model is equivalent to Model 6 [17]. We build on this work by providing a systematic Bayesian framework for model comparison and parameter estimation.

While our analysis employed sequence-dependent stochastic models, comparisons can also be made with some deterministic models.

Abbondanzieri et al. 2005 [4], Larson et al. 2012 [41], Schweikhard et al. 2014 [26], and Thomen et al. 2008. [27, 37] described a deterministic model (for RNAP, pol II, pol II, and T7 pol respectively) which estimated k_{cat}, K_D and translocation equilibrium constant $K_\tau = \frac{k_{bck}}{k_{fwd}}$. These are most similar to Model 4.

Maoiléidigh et al. 2011 for RNAP, and Dangkulwanich et al. 2013 for pol II, however found that the translocation and catalysis were occurring on similar timescales, and modelled only NTP binding as an equilibrium process [17, 42]. They also estimated the distance of translocation. These deterministic models are most similar to Model 11.

Finally, Mejia et al. 2015 [43] used a model that is quite different to all the above models, as it does not explicitly treat translocation. Instead elongation is modelled with a two step kinetic scheme, the first step involving NTP binding and conformational change, and the second step involving nucleotide incorporation and product release. This model is most similar to a special case of Model 5 where $\Delta G_{\tau 1}$ becomes extremely negative, driving the polymerase into the post-translocated position.

Results and discussion

Model selection with MCMC-ABC

Our aim was to 1) use Bayesian inference to select the best of 12 transcription elongation models for each RNA polymerase; and 2) estimate the parameters for those models appearing in the 95% credible set of the posterior distribution. Selecting prior distributions behind each parameter is a critical process in Bayesian inference. A prior distribution should reflect what is known about the parameter before observing the new data. We have explicated our prior assumptions, with justifications, in Table 1.

We performed MCMC-ABC experiments which estimated the parameters and model indicator M_i for $i \in \mathbb{Z}, 1 \leq i \leq 12$. Models which appear more often in this posterior distribution are better choices, given the data. The model indicator is a discrete variable which can take 12 values, and is treated identically to the 6 continuous parameters in the Bayesian framework.

The datasets we fit our models to are all from the single-molecule literature and are presented in: Figures 5a and 5b of Abbondanzieri et al. 2005 [4] for *E. coli* RNAP, Figure 2a of Schweikhard et al. 2014 [26] for *S. cerevisiae* pol II, and Table 2 of Thomen et al. 2008 [27] for T7 pol. To computationally replicate these experiments as faithfully as we could with the available information and computational limitations, simulations in this study were run on the 4 kb *E. coli rpoB* gene for RNAP (GenBank: EU274658), the first 4.75 kb of the human *rpb1* gene for pol II (NCBI: NG_027747) the first 10 kb of the Enterobacteria phage λ genome for T7 pol (NCBI: NC_001416). The mean velocities from 32 (for RNAP), 10 (for pol II) and 3 (for T7 pol) simulations of the full respective sequences were used to estimate the mean elongation velocity during MCMC-ABC, given F and [NTP].

For further information about the MCMC-ABC algorithm [25, 44], or the model indicator M_i, see S3 Appendix.

The posterior distributions

The posterior distributions from our MCMC-ABC experiments are presented in Table 2, Figs 5 and 6.

A large effective sample size (> 100 [53]) and a small $\hat{R}$ (< 1.1, as defined by Gelman et al. 1992 [54–56]) are essential for making reliable parameter estimates. Table 2 suggests that the parameters in the 95% credible set of models are sufficiently estimated by these criteria.

Table 1. Prior distributions used during Bayesian inference.

Parameter	Prior distribution(s)	Justification of prior distribution(s)
Model M	$P(M_i) = 2/16$ for $i \in \{1, 2, 4, 5\}$ $P(M_i) = 1/16$ for $i \in \{3, 6, 7, 8, 9, 10, 11, 12\}$	Each model should each have uniformly distributed values. Models with translocation at equilibrium have double the prior probability since these models do not use δ_1.
k_{cat} (s^{-1})	Lognormal($\mu = 3.454$, $\sigma = 0.587$) **for RNAP/pol II** Lognormal($\mu = 4.585$, $\sigma = 0.457$) **for T7 pol**	k_{cat} and elongation velocity estimates for *E. coli* RNAP and *S. cerevisiae* pol II range from 18 to 50 s^{-1} for optical trapping experiments [6–8, 21, 43], but as much as 100 bp/s *in vivo* [45–48]. Distribution selected such that (10, 100) is central 95% interval. For T7 pol k_{cat} and elongation velocity estimates range from 43–240 bp/s [9, 49–51]. Distribution selected such that (40, 240) is central 95% interval.
K_D (μM)	Lognormal($\mu = 1.844$, $\sigma = 1.762$)	Estimates for K_D under binding equilibrium models range from 20-140 μM [6, 20, 37, 41, 52]. In models where binding is kinetic and slow, $K_D \equiv \frac{k_{rel}}{k_{bind}}$ could be much lower (S4 Appendix). To accommodate for both binding models, the prior distribution was selected such that the central 95% interval is (0.2, 200).
k_{bind} ($\mu M^{-1} s^{-1}$)	Lognormal($\mu = -1.498$, $S\sigma = 1.585$)	Central 95% interval set so that NTP binding is a slow kinetic step (S4 Appendix). Centered around (0.01, 5).
$\Delta G_{\tau 1}$ ($k_B T$)	Normal($\mu = 0$, $\sigma = 1.55$) **for RNAP/ pol II** Normal($\mu = -3.3$, $\sigma = 1.55$) **for T7 pol**	For RNAP and pol II, centered around 0 with a standard deviation comparable to the free energy of a single nucleotide basepair doublet, and such that the 95% central interval is (-4, 4). For T7 pol $\Delta G_{\tau 1}$ has been estimated as -4.3 [37] and -4.87 $k_B T$ [35]. However these estimates are likely resulting partially from dangling ends. Thus, we subtracted the mean dangling end contribution of $\sim$ -1 $k_B T$ [33] and centered the prior around this interval with a standard deviation the same as above.
$\Delta G_{\tau}^{\ddagger}$ ($k_B T$)	Normal($\mu = 5.5$, $\sigma = 0.97$) **for RNAP/pol II** Normal($\mu = 2.5$, $\sigma = 1.36$) **for T7 pol**	Central 95% interval set so that translocation is a slow kinetic step (S4 Appendix). Selected so that 99% central interval is (3, 8) for RNAP and pol II, and (-1, 6) for T7 pol.
δ_1 (Å)	Uniform($l = 0$, $u = 3.4$)	Uniformly distributed across all possible values.

Prior distributions behind all estimated parameters and the model indicator. Unless specified otherwise, the prior distribution is used for all three RNA polymerases. Lognormal priors (parameterised in log space) are used for rates and equilibrium constants while normal priors are used for Gibbs energy terms. To maintain statistical integrity of the Bayesian analysis, prior distributions were not derived from the data presented by Abbondanzieri et al. 2005 [4] for RNAP, by Schweikhard et al. 2014 [26] for pol II, or by Thomen et al. 2008 [27] for T7 pol.

https://doi.org/10.1371/journal.pcbi.1006717.t001

These results indicate that the best models for the datasets examined are Models 11 and 12 for both RNAP and pol II, and Model 5 for T7 pol (Fig 4B).

For pol II, Model 12 has the highest posterior probability $P(M_{12}|D) = 0.71$. This is the most complex model considered, with 6 estimated parameters. In Model 12 translocation, NTP binding and catalysis are all kinetic processes; the displacement required to facilitate formation of the translocation transition state, $\delta_1 < \delta$, is estimated ($\hat{\delta}_1 = 3.1$ Å); and the standard Gibbs energy of the posttranslocated state is influenced by parameter $\Delta G_{\tau 1} \neq 0$.

The posterior distribution for RNAP consists of the same set models as that of pol II. For RNAP, Model 11 has the highest probability $P(M_{11}|D) = 0.81$. This model is a submodel of Model 12 with one fewer parameter: in Model 11 NTP binding is treated as an equilibrium process while in Model 12 it is not.

The only model in the 95% credible set for T7 pol is Model 5 $P(M_5|D) = 0.96$. In Model 5 (4 parameters) translocation, but not binding, is treated as an equilibrium process, and $\Delta G_{\tau 1}$ is estimated. This positions T7 pol in a quite different area of the model space to the other two polymerases.

Translocation rates differ among RNA polymerases

For RNAP and pol II, we estimate that a partial equilibrium approximation for the translocation step is inadequate. The posterior probability that such models are inadequate is 1.00 (see Table 2). For T7 pol, however, translocation is significantly faster than catalysis and is best modelled with a partial equilibrium approximation. Using estimates for $\Delta G_{\tau}^{\ddagger}$ and $\Delta G_{\tau 1}$ under the maximum posterior models (Model 11 for RNAP and Model 12 for pol II) we estimate the

Table 2. Summary of MCMC-ABC experiments.

	Enzyme	*E. coli* RNAP		*S. cerevisiae* pol II		Bacteriophage T7 pol
	ϵ	2.39		0.705		4.63
	Number of chains	70		10		26
	Combined chain length	3.5×10^7		6.2×10^7		1.2×10^8
	i	11	12	11	12	5
Model	Description	Binding equilibrium, Translocation kinetic	Binding kinetic, Translocation kinetic	Binding equilibrium, Translocation kinetic	Binding kinetic, Translocation kinetic	Binding kinetic, Translocation equilibrium
ESS / $\hat{R}$	$\hat{k}_{cat}$	257 / 1.04	1441 / 1.03	549 / 1.02	1203 / 1.01	2110 / 1.00
	$\widehat{\frac{k_{rel}}{k_{bind}}}$	328 / 1.01	101 / 1.01	536 / 1.01	133 / 1.05	106 / 1.00
	$\hat{k}_{bind}$	–	705 / 1.09	–	516 / 1.02	154 / 1.00
	$\Delta\hat{G}_{\tau 1}$	466 / 1.02	1844 / 1.00	1145 / 1.00	2769 / 1.00	1626 / 1.02
	$\hat{\delta}_1$	300 / 1.04	2290 / 1.03	658 / 1.01	1469 / 1.00	–
	$\Delta\hat{G}^{\ddagger}_{\tau}$	340 / 1.02	1680 / 1.03	589 / 1.02	1179 / 1.00	–
Posterior	$P(M_i \mid D)$	0.81	0.19	0.29	0.71	0.96

Each column summarises the posterior distribution for the respective RNA polymerase, which arises from multiple independent MCMC chains. The combined chain length refers to the total number of states sampled, a small fraction of which are used to estimate the 6 continuous model parameters and the model indicator M. Approximate Bayesian computation threshold ϵ is shown for each enzyme; state Θ is accepted into the posterior distribution only if $X^2(\Theta) \leq \epsilon$ (S3 Appendix). Models which appear in an RNA polymerase's 95% credible set and their posterior probabilities $P(M_i|D)$ are shown. The effective sample size (ESS, calculated with Tracer 1.6 [53]) and R-hat ($\hat{R}$ [54–56]) of each parameter, conditional on M_i, are displayed. A large ESS (> 100) and a small $\hat{R}$ (< 1.1) imply that the MCMC experiment has converged. Where a parameter is not incorporated in the kinetic model, a '–' is left in its place.

https://doi.org/10.1371/journal.pcbi.1006717.t002

mean forward $\bar{k}_{fwd}$ and backward $\bar{k}_{bck}$ translocation rates averaged across the *rpoB* sequence as: 230 s^{-1} and 112 s^{-1} for RNAP, and 350 s^{-1} and 12.7 s^{-1} for pol II, respectively (3 sf). These estimates are within one order of magnitude of the respective estimate for the rate of catalysis (Fig 5) suggesting that translocation and catalysis indeed occur on similar timescales.

For RNAP and pol II, translocation has frequently been modelled as an equilibrium process [4, 21, 26, 41, 43], however in some recent analyses this assumption has been rejected [16, 17, 42, 57, 58]. Our Bayesian analysis supports this. In contrast, there is general agreement that translocation in T7 pol is adequately modelled as an equilibrium process [27, 59, 60].

The data does not determine the kinetics of the NTP binding step

It remains unclear how to best model the NTP binding step. Models that describe NTP binding as a kinetic process have posterior probabilities of 0.19 for RNAP, 0.71 for pol II and 0.96 for T7 pol (Table 2). However, in an earlier experiment, where we used different a prior distribution for $\frac{k_{rel}}{k_{bind}}$, the latter probability was 0.21 and $P(M_4|D)$ was 0.79. The intermediate magnitude of these posterior probabilities, and sensitivity to the choice of prior, imply that the data contains very little information about which binding model is preferred.

Furthermore, $\frac{k_{rel}}{k_{bind}}$ and k_{bind} (Models 5 and 12) are unable to be estimated simultaneously. For pol II and for T7 pol, k_{bind} is estimated at around 0.48 and 1.4 μM^{-1} s^{-1} respectively with fairly narrow 95% highest posterior density (HPD) intervals (Fig 5). However, the HPD interval of $\frac{k_{rel}}{k_{bind}}$ spans three orders of magnitude and the value of this parameter was therefore poorly informed by the data. For RNAP, in contrast, neither k_{bind} nor $\frac{k_{rel}}{k_{bind}}$ were well-informed by the

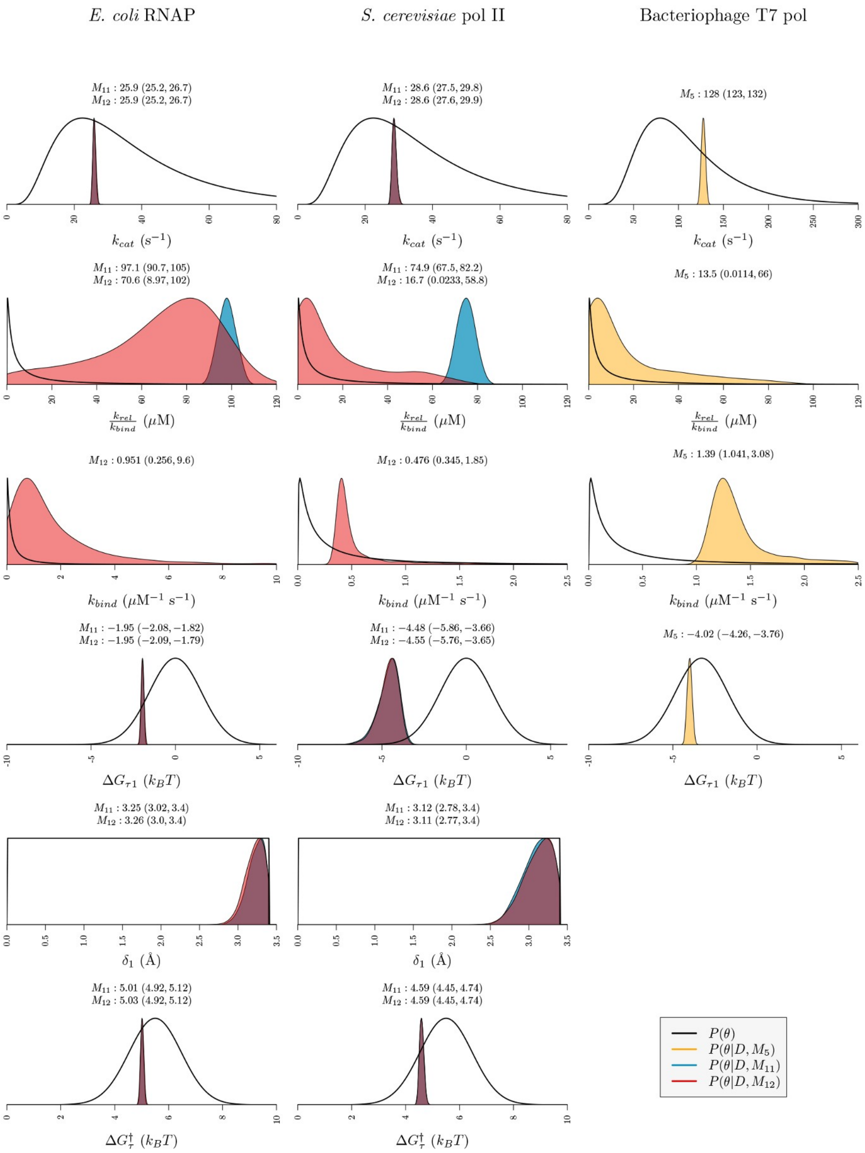

Fig 5. Posterior and prior distribution plots. Posterior distributions for all models which appear in the 95% credible set are displayed (two models for RNAP, two models for pol II, and one model for T7 pol). Plots show the prior probability density $P(\theta)$ of each parameter and posterior probability density of each parameter conditional on the model $P(\theta|D, M_i)$. The geometric median point-estimates and highest posterior density (HPD) intervals (calculated with Tracer 1.6 [53]) are displayed above each plot (3 sf).

https://doi.org/10.1371/journal.pcbi.1006717.g005

data and both have HPD intervals spanning 1-2 orders of magnitude. This non-identifiability—where two or more parameters are unable to be estimated simultaneously (S4 Appendix)—highlights the appeal of an NTP binding equilibrium model where only one parameter $\frac{k_{rel}}{k_{bind}}$ needs to be estimated, despite the unrealistic assumptions it may invoke. In the case of each enzyme, the data has taught us nothing about one or two of the binding parameters.

The pause-free mean velocities measured during transcription elongation follow Michaelis-Menten kinetics even though the reaction cycle is more complicated than that of a simple enzyme [61]. As such, the inability to resolve the timescale of the substrate binding step is unsurprising [62–64].

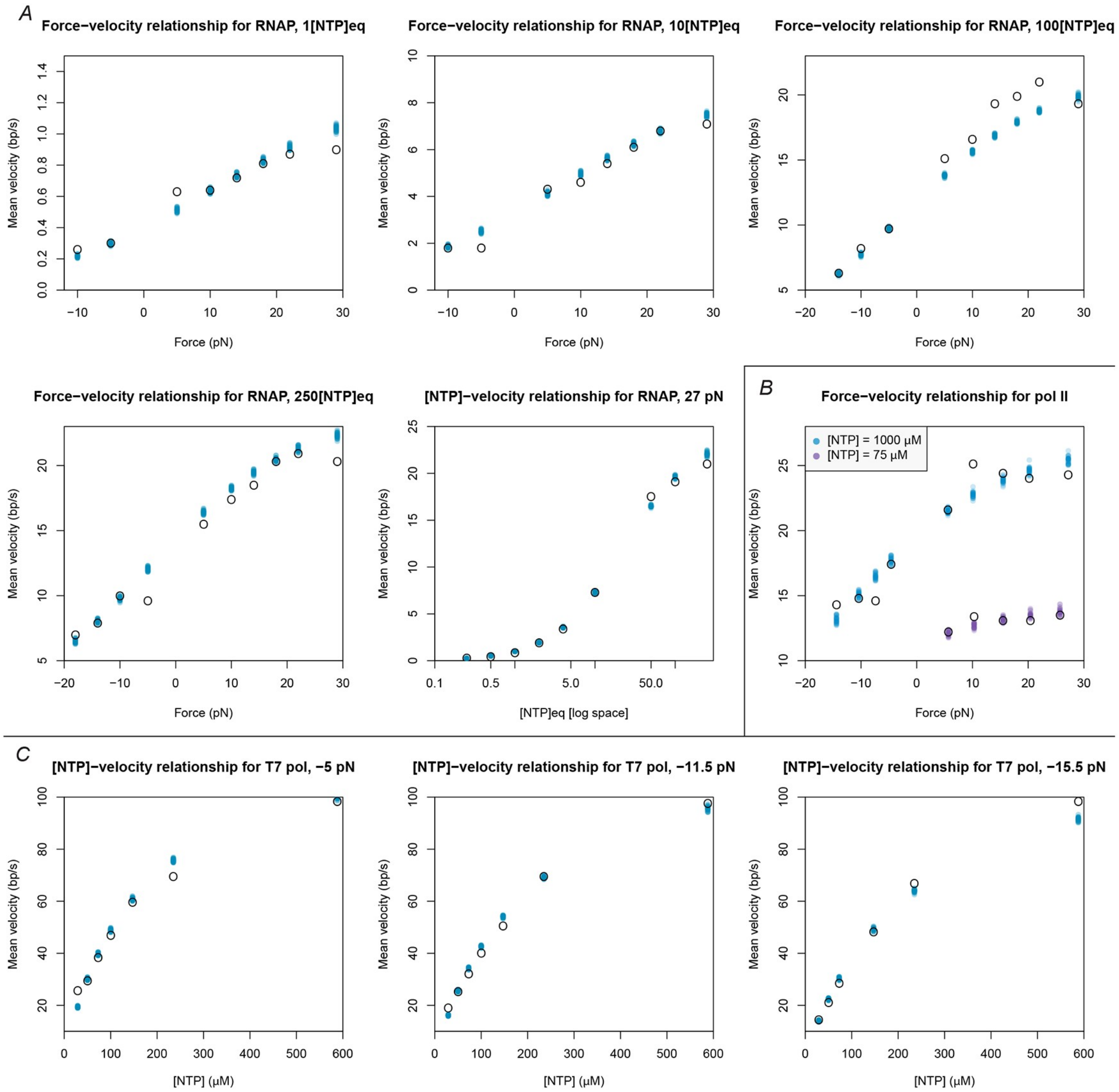

Fig 6. Posterior distributions of simulated velocities. Black open circles represent experimentally measured mean velocities reported in the original publication for (A) RNAP, (B) pol II, and (C) T7 pol [4, 26, 27]. Each coloured dot represents a single sample simulated from the posterior distribution of parameters/models for the respective polymerase. 30 samples were generated from each of the three posterior distributions. For RNAP, $[NTP]_{eq}$ is defined as $[ATP] = 5\ \mu M$, $[CTP] = 2.5\ \mu M$, $[GTP] = 10\ \mu M$, and $[UTP] = 10\ \mu M$.

https://doi.org/10.1371/journal.pcbi.1006717.g006

In the transcription literature, NTP binding is almost always assumed to achieve equilibrium for RNAP, pol II, and T7 pol [4, 16, 17, 21, 26, 27, 37, 41, 42, 60]. However Mejia et al. 2015 [43] have shown that NTP binding is indeed rate-limiting, and that mutations in the RNAP trigger loop impair the binding rate thus suggesting that the trigger loop is coupled with NTP binding.

RNAP has an energetic preference for the posttranslocated state

In previous stochastic sequence-dependent models [16, 21] the standard Gibbs energies of the pre and posttranslocated states have been based solely on the nucleic acid basepairing energies. Our models include an additional term, $\Delta G_{\tau 1}$, to account for potential interactions between the protein and the nucleic acid. The marginal posterior probability of a model in which an additional term $\Delta G_{\tau 1}$ is required is 1.00 in all three polymerases. In each case $\Delta G_{\tau 1}$ was estimated to be less than 0 $k_B T$ and 0 $k_B T$ is not included in the 95% HPD interval (Fig 5). We find that $\Delta \hat{G_{\tau 1}}$ is the most significant in pol II and T7 pol: −4.6 $k_B T$ and −4.0 $k_B T$ respectively, while $\Delta \hat{G_{\tau 1}} = -2.0\ k_B T$ for RNAP (2 sf).

These results suggest that structural elements within RNA polymerases can energetically favour posttranslocated states over pretranslocated states. We note that the sequence-dependent contribution of the dangling end of the DNA/RNA hybrid is included in the thermodynamic model. The energetic bias for the posttranslocated state is separable from this effect.

To facilitate comparison with previous deterministic models, using our estimates of $\Delta G_{\tau 1}$ we calculated the equilibrium constant between the pre and posttranslocated states. Geometrically averaged across the *rpoB* gene, these are

$$\bar{K}_{\tau} = \frac{1}{L - l_0} exp\{\sum_{l=l_0}^{L-1} \ln(k_{bck}(l)/k_{fwd}(l))\} = \begin{cases} 0.77 \text{ for RNAP} \\ 0.057 \text{ for pol II} \\ 0.10 \text{ for T7 pol.} \end{cases} \tag{11}$$

Thus, for all three polymerases, $K_\tau < 1$, indicating that the small energetic preference that the protein has for the posttranslocated state is sufficient to override the loss of basepairing energy, thereby biasing the system towards population of the posttranslocated positions. This is in agreement with estimates made for pol II and T7 pol [26, 27, 35, 36, 41] and Kireeva et al. 2018 [58] for RNAP: *"forward translocation occurs in milliseconds and is poorly reversible"*. However these estimates are inconsistent with some RNAP and pol II studies which place this ratio above 1 [4, 17, 42, 52].

Kinetic modelling can itself suggest no physical mechanism for the stabilisation. Yu et al. 2012 [36] have identified a conserved tyrosine residue near the active site of T7 pol that pushes against the 3′ end of the mRNA, and thus stabilises the posttranslocated state. They propose a similar mechanism for the multi-subunit RNA polymerases.

δ_1 may be an important parameter but its physical meaning is unclear

Our results suggest that δ_1, the distance that RNA polymerase must translocate forward by to reach the translocation transition state, is a necessary parameter to estimate for RNAP and pol II. Setting $\delta_1 = \delta/2$ is not sufficient. The marginal posterior probability of models which estimate this term is 1.00. δ_1 is irrelevant to the modeling of the T7 pol data because the best models invoke a partial equilibrium approximation for the translocation step.

While our prior distribution restricted δ_1 to lie in the range (0, δ), the upper end our 95% HPD intervals of δ_1 for RNAP and pol II are very close to δ = 3.4 Å. If it was not for this prior distribution, δ_1 estimates would have included values higher than δ. Similar results have been observed by Maoiléidigh et al. 2011 [17] for RNAP.

Our interpretation of δ_1 implies it should never be greater than δ nor should δ be more than the width of one basepair. The physical meaning of δ_1 with values greater than δ is thus unclear. It is noted that δ_1 is only used when $F \neq 0$.

Comparing the kinetics of RNA polymerases

The *in vivo* rate of transcription elongation varies considerably across RNAP, pol II and T7 pol. The prokaryotic and eukaryotic RNA polymerases have a mean rate ranging from 20-120 bp/s [45, 46, 48, 49, 65–67], which may be slowed down in histone-wrapped regions of eukaryotic genomes [7]. In contrast, Bacteriophage T7 pol operates up to an order of magnitude faster (around 200-240 bp/s [49, 68]) and is known to be quite insensitive to transcriptional pause sites [9, 27].

In additional to these differences, we have shown that translocation is very rapid in T7 pol, relative to the rate of NTP incorporation, while the disparity is much less significant in RNAP and pol II. Furthermore, the model does not fit the data for T7 pol as closely it does for RNAP and pol II (Fig 6). T7 pol therefore seems to operate under quite a different kinetic scheme than that of the cellular polymerases, which is not unexpected given their distant evolutionary relationship [3].

In general, the elongation velocity of RNA polymerase is significantly slower in an optical trap (with estimates ranging from 9.7-22 bp/s for RNAP [11–13, 43, 69]) compared with that of the untethered enzyme (with estimates *in vitro* or *in vivo* ranging from 25-118 bp/s for RNAP [45, 49, 70, 71]). This relationship holds for multiple RNA polymerases including *E. coli* RNAP, *S. cerevisiae* pol II [41, 42, 52, 72], Bacteriophage T7 pol [9, 27, 49, 51], and Bacteriophage Φ6 P2 [10, 73]. This suggests that optical trapping perturbs the system to a significant extent. Additionally, varying degrees of heterogeneity in elongation rate have been observed across different polymerase complexes even under the same conditions [11, 13, 27].

The velocity perturbations resulting from the optical trapping apparatus will be propagated into the model parameters, especially k_{cat}, and $\Delta G_{\tau}^{\ddagger}$, and some caution is needed when extrapolating these results to untethered systems.

Bayesian inference of transcription elongation

To our knowledge we are the first to perform Bayesian inference on single-molecule models of transcription elongation. This was achieved by simulation which necessitated the use of approximate Bayesian computation. An alternative would be to build and use a likelihood function (ie. the probability of taking exactly t units of time for RNA polymerase to copy the sequence n times). The latter approach can be achieved using chemical master equations, as opposed to (Gillespie) sampling from the distribution. Finding analytical, stable numerical, or approximate solutions to the chemical master equations could provide a similar insight in less computational time, however is susceptible to a multitude of analytical and numerical issues associated with the exponentiation of an arbitrary transition rate matrix that grows with the length of the sequence (S2 Appendix) [74]. This problem would be amplified by the introduction of backtracking, hypertranslocation, or NTP misincorporation reactions into the model, for instance. The Bayesian framework we have presented, although computationally intensive due to its simulation requirement, is general and will work on any model of transcription without the need to resolve these issues. The path has been paved for modelling transcriptional pausing, for instance [16, 21, 75]. Nevertheless, likelihood-based Bayesian inference is an approach that should be explored in the future.

We have demonstrated that single-molecule data can be usefully analysed using a Bayesian inference and model selection framework. This analysis would have even greater statistical power if applied to the progression of individual RNA polymerase complexes instead of mean velocities averaged across multiple experiments.

Conclusion

In this article we evaluated some simple Brownian ratchet models of transcription elongation (Fig 2). By varying the parameterisation of the translocation step (Fig 3) and incorporating partial equilibrium approximations commonly invoked in the literature (Fig 4A) we enumerated a total of 12 related models (Fig 4B). Using stochastic simulations and approximate

Bayesian computation, we then assessed which of these models were capable of describing the force-velocity data previously measured for several RNA polymerases (Table 2 and Fig 5) using single-molecule optical trapping experiments [4, 26, 27].

Our analysis suggests that 1) different partial equilibrium approximations of the translocation step are appropriate for the multisubunit RNA polymerases versus the single subunit T7 RNA polymerase. 2) Treatment of the NTP binding step remains a point of ambiguity. The existing data does not place strong constraints on the modelling of this step. 3) There is an energetic bias for posttranslocated state. 4) The model of the force-dependent translocation, which invokes transition state theory, is not physically realistic.

Supporting information

S1 Appendix. Stochastic simulation. The Gillespie algorithm is described.
(PDF)

S2 Appendix. Chemical master equations. Master equations for the four equilibrium model variants are presented.
(PDF)

S3 Appendix. MCMC-ABC. A description of the MCMC-ABC algorithm and how it is used to infer parameters and models from experimental data.
(PDF)

S4 Appendix. Prior distributions. Simulation-based justifications behind prior distributions for $\Delta G^{\ddagger}_{\tau}$, k_{bind}, and $\frac{k_{rel}}{k_{bind}}$.
(PDF)

S1 Fig. Simulations of the elongation pathway. Each point is a single simulation of the full *rpoB* gene (4029 nt). For (A-C), Parameters on the x- and z-axis are sampled uniformly at random from the displayed range at the beginning of each trial. The y-axis of each plot (mean elongation velocity) is then measured from the respective simulation. [NTP] and F held constant at 1000 μM and 0 pN respectively. (A) and (B): Relationship between $\Delta G^{\ddagger}_{\tau}$ and k_{cat} for the melting model with binding at equilibrium (Model 8). $\Delta G_{\tau 1}$ set to its prior mean (0 for RNAP and pol II, and -3.3 for T7 pol). (C) Relationship between k_{bind} and k_{cat} for the kinetic binding model with translocation at equilibrium (Model 2). (D) Relationship between K_D and k_{bind} with translocation held at equilibrium (Model 2). K_D and k_{bind} sampled uniformly from specified range and velocity is measured. Samples with simulated velocities outside of the range 1-2 bp/s were discarded. [NTP] = 10 μM and k_{cat} = 100 s^{-1}.
(PDF)

Author Contributions

Conceptualization: Jordan Douglas.

Data curation: Jordan Douglas.

Formal analysis: Jordan Douglas.

Investigation: Jordan Douglas, Richard Kingston, Alexei J. Drummond.

Methodology: Jordan Douglas, Richard Kingston, Alexei J. Drummond.

Project administration: Richard Kingston, Alexei J. Drummond.

Resources: Richard Kingston, Alexei J. Drummond.

Software: Jordan Douglas.

Supervision: Richard Kingston, Alexei J. Drummond.

Validation: Jordan Douglas.

Visualization: Jordan Douglas.

Writing – original draft: Jordan Douglas.

Writing – review & editing: Jordan Douglas, Richard Kingston, Alexei J. Drummond.

References

1. Sweetser D, Nonet M, Young RA. Prokaryotic and eukaryotic RNA polymerases have homologous core subunits. Proceedings of the National Academy of Sciences. 1987; 84(5):1192–1196. https://doi.org/10.1073/pnas.84.5.1192
2. Sosunov V, Sosunova E, Mustaev A, Bass I, Nikiforov V, Goldfarb A. Unified two-metal mechanism of RNA synthesis and degradation by RNA polymerase. The EMBO journal. 2003; 22(9):2234–2244. https://doi.org/10.1093/emboj/cdg193 PMID: 12727889
3. Sousa R, Chung YJ, Rose JP, Wang BC. Crystal structure of bacteriophage T7 RNA polymerase at 3.3 Å resolution. Nature. 1993; 364(6438):593. https://doi.org/10.1038/364593a0 PMID: 7688864
4. Abbondanzieri EA, Greenleaf WJ, Shaevitz JW, Landick R, Block SM. Direct observation of base-pair stepping by RNA polymerase. Nature. 2005; 438(7067):460–465. https://doi.org/10.1038/nature04268 PMID: 16284617
5. Adelman K, La Porta A, Santangelo TJ, Lis JT, Roberts JW, Wang MD. Single molecule analysis of RNA polymerase elongation reveals uniform kinetic behavior. Proceedings of the National Academy of Sciences. 2002; 99(21):13538–13543. https://doi.org/10.1073/pnas.212358999
6. Bai L, Fulbright RM, Wang MD. Mechanochemical kinetics of transcription elongation. Physical review letters. 2007; 98(6):068103. https://doi.org/10.1103/PhysRevLett.98.068103 PMID: 17358986
7. Hodges C, Bintu L, Lubkowska L, Kashlev M, Bustamante C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. Science. 2009; 325(5940):626–628. https://doi.org/10.1126/science.1172926 PMID: 19644123
8. Galburt EA, Grill SW, Wiedmann A, Lubkowska L, Choy J, Nogales E, et al. Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner. Nature. 2007; 446(7137):820–823. https://doi.org/10.1038/nature05701 PMID: 17361130
9. Skinner GM, Baumann CG, Quinn DM, Molloy JE, Hoggett JG. Promoter binding, initiation, and elongation by bacteriophage T7 RNA polymerase a single-molecule view of the transcription cycle. Journal of Biological Chemistry. 2004; 279(5):3239–3244. https://doi.org/10.1074/jbc.M310471200 PMID: 14597619
10. Dulin D, Vilfan ID, Berghuis BA, Hage S, Bamford DH, Poranen MM, et al. Elongation-competent pauses govern the fidelity of a viral RNA-dependent RNA polymerase. Cell reports. 2015; 10(6):983–992. https://doi.org/10.1016/j.celrep.2015.01.031 PMID: 25683720
11. Neuman KC, Abbondanzieri EA, Landick R, Gelles J, Block SM. Ubiquitous transcriptional pausing is independent of RNA polymerase backtracking. Cell. 2003; 115(4):437–447. https://doi.org/10.1016/s0092-8674(03)00845-6 PMID: 14622598
12. Davenport RJ, Wuite GJ, Landick R, Bustamante C. Single-molecule study of transcriptional pausing and arrest by E. coli RNA polymerase. Science. 2000; 287(5462):2497. https://doi.org/10.1126/science.287.5462.2497 PMID: 10741971
13. Tolić-Nørrelykke SF, Engh AM, Landick R, Gelles J. Diversity in the rates of transcript elongation by single RNA polymerase molecules. Journal of Biological Chemistry. 2004; 279(5):3292–3299. https://doi.org/10.1074/jbc.M310290200 PMID: 14604986
14. Abbondanzieri EA, Shaevitz JW, Block SM. Picocalorimetry of transcription by RNA polymerase. Biophysical journal. 2005; 89(6):L61–L63. https://doi.org/10.1529/biophysj.105.074195 PMID: 16239336
15. Watson JD, Crick FH, et al. Molecular structure of nucleic acids. Nature. 1953; 171(4356):737–738. https://doi.org/10.1038/171737a0 PMID: 13054692
16. Tadigotla VR, Maoiléidigh DÓ, Sengupta AM, Epshtein V, Ebright RH, Nudler E, et al. Thermodynamic and kinetic modeling of transcriptional pausing. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(12):4439–4444. https://doi.org/10.1073/pnas.0600508103 PMID: 16537373

17. Maoiléidigh DÓ, Tadigotla VR, Nudler E, Ruckenstein AE. A unified model of transcription elongation: what have we learned from single-molecule experiments? Biophysical journal. 2011; 100(5):1157–1166. https://doi.org/10.1016/j.bpj.2010.12.3734

18. Maitra U, Nakata Y, Hurwitz J. The Role of Deoxyribonucleic Acid in Ribonucleic Acid Synthesis XIV. A Study of the Initiation of Ribonucleic Acid Synthesis. Journal of Biological Chemistry. 1967; 242 (21):4908–4918. PMID: 4862425

19. Erie DA, Yager TD, Von Hippel PH. The single-nucleotide addition cycle in transcription: a biophysical and biochemical perspective. Annual review of biophysics and biomolecular structure. 1992; 21 (1):379–415. https://doi.org/10.1146/annurev.bb.21.060192.002115 PMID: 1381976

20. Rhodes G, Chamberlin MJ. Ribonucleic acid chain elongation by Escherichia coli ribonucleic acid polymerase I. Isolation of ternary complexes and the kinetics of elongation. Journal of Biological Chemistry. 1974; 249(20):6675–6683. PMID: 4608711

21. Bai L, Shundrovsky A, Wang MD. Sequence-dependent kinetic model for transcription elongation by RNA polymerase. Journal of molecular biology. 2004; 344(2):335–349. https://doi.org/10.1016/j.jmb.2004.08.107 PMID: 15522289

22. Bustamante C, Chemla YR, Forde NR, Izhaky D. Mechanical processes in biochemistry. Annual review of biochemistry. 2004; 73(1):705–748. https://doi.org/10.1146/annurev.biochem.72.121801.161542 PMID: 15189157

23. Cleland W. Partition analysis and concept of net rate constants as tools in enzyme kinetics. Biochemistry. 1975; 14(14):3220–3224. https://doi.org/10.1021/bi00685a029 PMID: 1148201

24. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry. 1977; 81(25):2340–2361. https://doi.org/10.1021/j100540a008

25. Beaumont MA. Approximate Bayesian computation in evolution and ecology. Annual review of ecology, evolution, and systematics. 2010; 41:379–406. https://doi.org/10.1146/annurev-ecolsys-102209-144621

26. Schweikhard V, Meng C, Murakami K, Kaplan CD, Kornberg RD, Block SM. Transcription factors TFIIF and TFIIS promote transcript elongation by RNA polymerase II by synergistic and independent mechanisms. Proceedings of the National Academy of Sciences. 2014; 111(18):6642–6647. https://doi.org/10.1073/pnas.1405181111

27. Thomen P, Lopez P, Bockelmann U, Guillerez J, Dreyfus M, Heslot F. T7 RNA polymerase studied by force measurements varying cofactor concentration. Biophysical journal. 2008; 95(5):2423–2433. https://doi.org/10.1529/biophysj.107.125096 PMID: 18708471

28. Wang MD, Schnitzer MJ, Yin H, Landick R, Gelles J, Block SM. Force and velocity measured for single molecules of RNA polymerase. Science. 1998; 282(5390):902–907. https://doi.org/10.1126/science.282.5390.902 PMID: 9794753

29. Shaevitz JW, Abbondanzieri EA, Landick R, Block SM. Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. Nature. 2003; 426(6967):684–687. https://doi.org/10.1038/nature02191 PMID: 14634670

30. Artsimovitch I, Landick R. Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. Proceedings of the National Academy of Sciences. 2000; 97(13):7090–7095. https://doi.org/10.1073/pnas.97.13.7090

31. Zhou Y, Navaroli DM, Enuameh MS, Martin CT. Dissociation of halted T7 RNA polymerase elongation complexes proceeds via a forward-translocation mechanism. Proceedings of the National Academy of Sciences. 2007; 104(25):10352–10357. https://doi.org/10.1073/pnas.0606306104

32. Greive SJ, Von Hippel PH. Thinking quantitatively about transcriptional regulation. Nature Reviews Molecular Cell Biology. 2005; 6(3):221–232. https://doi.org/10.1038/nrm1588 PMID: 15714199

33. SantaLucia J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proceedings of the National Academy of Sciences. 1998; 95(4):1460–1465. https://doi.org/10.1073/pnas.95.4.1460

34. Wu P, Nakano Si, Sugimoto N. Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation. The FEBS Journal. 2002; 269(12):2821–2830.

35. Yin YW, Steitz TA. The structural mechanism of translocation and helicase activity in T7 RNA polymerase. Cell. 2004; 116(3):393–404. https://doi.org/10.1016/s0092-8674(04)00120-5 PMID: 15016374

36. Yu J, Oster G. A small post-translocation energy bias aids nucleotide selection in T7 RNA polymerase transcription. Biophysical journal. 2012; 102(3):532–541. https://doi.org/10.1016/j.bpj.2011.12.028 PMID: 22325276

37. Thomen P, Lopez PJ, Heslot F. Unravelling the mechanism of RNA-polymerase forward motion by using mechanical force. Physical Review Letters. 2005; 94(12):128102. https://doi.org/10.1103/PhysRevLett.94.128102 PMID: 15903965

38. Depken M, Galburt EA, Grill SW. The origin of short transcriptional pauses. Biophysical journal. 2009; 96(6):2189–2193. https://doi.org/10.1016/j.bpj.2008.12.3918 PMID: 19289045

39. Herbert KM, Greenleaf WJ, Block SM. Single-molecule studies of RNA polymerase: motoring along. Annu Rev Biochem. 2008; 77:149–176. https://doi.org/10.1146/annurev.biochem.77.073106.100741 PMID: 18410247

40. Lecca P. Stochastic chemical kinetics. Biophysical reviews. 2013; 5(4):323–345. https://doi.org/10.1007/s12551-013-0122-2 PMID: 28510113

41. Larson MH, Zhou J, Kaplan CD, Palangat M, Kornberg RD, Landick R, et al. Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. Proceedings of the National Academy of Sciences. 2012; 109(17):6555–6560. https://doi.org/10.1073/pnas.1200939109

42. Dangkulwanich M, Ishibashi T, Liu S, Kireeva ML, Lubkowska L, Kashlev M, et al. Complete dissection of transcription elongation reveals slow translocation of RNA polymerase II in a linear ratchet mechanism. Elife. 2013; 2:e00971. https://doi.org/10.7554/eLife.00971 PMID: 24066225

43. Mejia YX, Nudler E, Bustamante C. Trigger loop folding determines transcription rate of Escherichia coli's RNA polymerase. Proceedings of the National Academy of Sciences. 2015; 112(3):743–748. https://doi.org/10.1073/pnas.1421067112

44. Csilléry K, Blum MG, Gaggiotti OE, François O. Approximate Bayesian computation (ABC) in practice. Trends in ecology & evolution. 2010; 25(7):410–418. https://doi.org/10.1016/j.tree.2010.04.001

45. Vogel U, Jensen KF. The RNA chain elongation rate in Escherichia coli depends on the growth rate. Journal of bacteriology. 1994; 176(10):2807–2813. https://doi.org/10.1128/jb.176.10.2807-2813.1994 PMID: 7514589

46. Ryals J, Little R, Bremer H. Temperature dependence of RNA synthesis parameters in Escherichia coli. Journal of bacteriology. 1982; 151(2):879–887. https://doi.org/10.1128/JB.151.2.879-887.1982 PMID: 6178724

47. Richardson JP, Greenblatt J. Control of RNA chain elongation and termination. Escherichia coli and Salmonella: cellular and molecular biology. 1996; 1:822–848.

48. Mason PB, Struhl K. Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. Molecular cell. 2005; 17(6):831–840. https://doi.org/10.1016/j.molcel.2005.02.017 PMID: 15780939

49. Iost I, Guillerez J, Dreyfus M. Bacteriophage T7 RNA polymerase travels far ahead of ribosomes in vivo. Journal of bacteriology. 1992; 174(2):619–622. https://doi.org/10.1128/jb.174.2.619-622.1992 PMID: 1729251

50. Bonner G, Lafer EM, Sousa R. Characterization of a set of T7 RNA polymerase active site mutants. Journal of Biological Chemistry. 1994; 269(40):25120–25128. PMID: 7929200

51. Anand VS, Patel SS. Transient state kinetics of transcription elongation by T7 RNA polymerase. Journal of Biological Chemistry. 2006; 281(47):35677–35685. https://doi.org/10.1074/jbc.M608180200 PMID: 17005565

52. Kireeva ML, Nedialkov YA, Cremona GH, Purtov YA, Lubkowska L, Malagon F, et al. Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation. Molecular cell. 2008; 30(5):557–566. https://doi.org/10.1016/j.molcel.2008.04.017 PMID: 18538654

53. Rambaut A, Drummond A. Tracer 1.6. University of Edinburgh, Edinburgh. UK. Technical report; 2013.

54. Gelman A, Rubin DB, et al. Inference from iterative simulation using multiple sequences. Statistical science. 1992; 7(4):457–472. https://doi.org/10.1214/ss/1177011136

55. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics. 1998; 7(4):434–455. https://doi.org/10.1080/10618600.1998.10474787

56. Brooks S, Gelman A, Jones G, Meng XL. Handbook of markov chain monte carlo. CRC press; 2011.

57. Nedialkov YA, Nudler E, Burton ZF. RNA polymerase stalls in a post-translocated register and can hyper-translocate. Transcription. 2012; 3(5):260–269. https://doi.org/10.4161/trns.22307 PMID: 23132506

58. Kireeva M, Trang C, Matevosyan G, Turek-Herman J, Chasov V, Lubkowska L, et al. RNA–DNA and DNA–DNA base-pairing at the upstream edge of the transcription bubble regulate translocation of RNA polymerase and transcription rate. Nucleic acids research. 2018; 46(11):5764–5775. https://doi.org/10.1093/nar/gky393 PMID: 29771376

59. Guajardo R, Lopez P, Dreyfus M, Sousa R. NTP concentration effects on initial transcription by T7 RNAP indicate that translocation occurs through passive sliding and reveal that divergent promoters have distinct NTP concentration requirements for productive initiation. Journal of molecular biology. 1998; 281(5):777–792. https://doi.org/10.1006/jmbi.1998.1988 PMID: 9719634

60. Arnold S, Siemann M, Scharnweber K, Werner M, Baumann S, Reuss M, et al. Kinetic modeling and simulation of in vitro transcription by phage T 7 RNA polymerase. Biotechnology and bioengineering. 2001; 72(5):548–561. https://doi.org/10.1002/1097-0290(20010305)72:5%3C548::AID-BIT1019%3E3.0.CO;2-2 PMID: 11460245

61. Wong F, Dutta A, Chowdhury D, Gunawardena J. Structural conditions on complex networks for the Michaelis–Menten input–output response. Proceedings of the National Academy of Sciences. 2018; 115(39):9738–9743. https://doi.org/10.1073/pnas.1808053115

62. Briggs GE, Haldane JBS. A note on the kinetics of enzyme action. Biochemical journal. 1925; 19(2):338. https://doi.org/10.1042/bj0190338 PMID: 16743508

63. English BP, Min W, Van Oijen AM, Lee KT, Luo G, Sun H, et al. Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. Nature chemical biology. 2006; 2(2):87–94. https://doi.org/10.1038/nchembio759 PMID: 16415859

64. Schnell S. Validity of the Michaelis–Menten equation–steady-state or reactant stationary assumption: that is the question. The FEBS journal. 2014; 281(2):464–472. https://doi.org/10.1111/febs.12564 PMID: 24245583

65. Tennyson CN, Klamut HJ, Worton RG. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. Nature genetics. 1995; 9(2):184–190. https://doi.org/10.1038/ng0295-184 PMID: 7719347

66. Darzacq X, Shav-Tal Y, De Turris V, Brody Y, Shenoy SM, Phair RD, et al. In vivo dynamics of RNA polymerase II transcription. Nature structural & molecular biology. 2007; 14(9):796–806. https://doi.org/10.1038/nsmb1280

67. Kainov DE, Lísal J, Bamford DH, Tuma R. Packaging motor from double-stranded RNA bacteriophage ϕ12 acts as an obligatory passive conduit during transcription. Nucleic acids research. 2004; 32 (12):3515–3521. https://doi.org/10.1093/nar/gkh680 PMID: 15247341

68. Makarova OV, Makarov EM, Sousa R, Dreyfus M. Transcribing of Escherichia coli genes with mutant T7 RNA polymerases: stability of lacZ mRNA inversely correlates with polymerase speed. Proceedings of the National Academy of Sciences. 1995; 92(26):12250–12254. https://doi.org/10.1073/pnas.92.26.12250

69. Mejia YX, Mao H, Forde NR, Bustamante C. Thermal probing of E. coli RNA polymerase off-pathway mechanisms. Journal of molecular biology. 2008; 382(3):628–637. https://doi.org/10.1016/j.jmb.2008.06.079 PMID: 18647607

70. Burns CM, Richardson LV, Richardson JP. Combinatorial effects of NusA and NusG on transcription elongation and rho-dependent termination in Escherichia coli1. Journal of molecular biology. 1998; 278 (2):307–316. https://doi.org/10.1006/jmbi.1998.1691 PMID: 9571053

71. Kingston R, Nierman W, Chamberlin M. A direct effect of guanosine tetraphosphate on pausing of Escherichia coli RNA polymerase during RNA chain elongation. Journal of Biological Chemistry. 1981; 256(6):2787–2797. PMID: 7009598

72. Galburt EA, Grill SW, Bustamante C. Single molecule transcription elongation. Methods. 2009; 48(4):323–332. https://doi.org/10.1016/j.ymeth.2009.04.021 PMID: 19426807

73. Usala SJ, Brownstein BH, Haselkorn R. Displacement of parental RNA strands during in vitro transcription by bacteriophage φ6 nucleocapsids. Cell. 1980; 19(4):855–862. https://doi.org/10.1016/0092-8674(80)90076-8 PMID: 7379123

74. Moler C, Van Loan C. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM review. 2003; 45(1):3–49. https://doi.org/10.1137/S00361445024180

75. Bai L, Wang MD. Comparison of pause predictions of two sequence-dependent transcription models. Journal of Statistical Mechanics: Theory and Experiment. 2010; 2010(12):P12007. https://doi.org/10.1088/1742-5468/2010/12/P12007